Chapter 7

Estimation with confidence

In this chapter the idea of using samples of data to provide estimates for unknown model parameter values is further developed. However, instead of stating a single number (a point estimate), the aim is to provide a range of plausible values for the parameters.

In this chapter the ideas of estimation which were introduced and described in Chapter 6 are extended. In Chapter 6, Example 6.2 an experiment in which 103 samples of water were collected from the same source was described. The number of specimens of the leech *Helobdella* in each sample was counted. A Poisson model was suggested for the variation observed in the counts. The data collected during the experiment yielded an estimate (the sample mean) $\bar{x} = \mu = 0.816$ for the Poisson parameter $\mu$, that is, for the underlying average number of leeches per water sample.

It was pointed out that had a second experiment been performed using the same water source, then the numbers of leeches collected would probably have been different, and the sample mean used to estimate the underlying mean $\mu$ would probably have been different. It has already been remarked several times that in a random sample the observed sample mean $\bar{x}$ is just one observation on a random variable $X$. Quite conceivably, in this experiment, $\bar{x}$ might have been as low as 0.7 or even lower; it might have been as high as 1, or perhaps 2, or 3. In the initial experiment (the one actually performed), there were a lot of 0s observed (samples containing no leeches at all), quite a few 1s and 2s, and just a few 3s and 4s. None of the 103 samples contained more than eight leeches. It would therefore be very surprising to find that the underlying mean incidence of leeches was as high as 6, say, and it is almost incredible (though, of course, it is not impossible) that the underlying mean could be as high as 8, or higher still.

This chapter is about using the results of statistical experiments to obtain some idea of a plausible range of values for some unknown population characteristic (maybe an average, or a rate, or a proportion). Assuming a reasonable statistical model, this characteristic will always be expressible in terms of the parameters of the model. The minimum and maximum values of this range are called confidence limits, and the range of plausible or 'believable' values is called a confidence interval.

Section 7.1 of the chapter deals with the situation where only a single observation has been collected in order to shed light on the area of investigation: our random sample is of size one. You will see how confidence intervals for the indexing parameters of some of the standard probability models with which
you are already familiar are calculated. You already know from Chapter 4 how important it is in a sampling context, generally speaking, to collect as much data as possible, and in this chapter the influence of sample size on the usefulness of the conclusions drawn will become very apparent. However, circumstances do occur where a sample is necessarily small, and this approach (drawing a sample of size one) provides an easy introduction to the topic.

In Section 7.2 you will see the consequences of increasing the sample size. It is very useful to be able to say with some high degree of confidence that the value of some unknown parameter is between certain limits. It is more useful still if those limits are not very far apart. Specifically, the confidence limits obtained through a sampling experiment have a tendency to move closer and closer together as the sample size increases: the resulting confidence interval becomes narrower and more ‘precise’. All the calculations in this section require a computer. (The problem is simple enough to state, but the arithmetic involved in calculating confidence limits can become surprisingly difficult.)

In Sections 7.1 and 7.2 we look at confidence intervals for descriptive parameters such as a Poisson mean $\mu$, a Bernoulli probability $p$, an exponential mean, and so on. Sections 7.3 and 7.4 are based on normal theory. In Section 7.3, it is assumed that interest centres around a population where the variation observed may be adequately modelled by a normal distribution. We have seen many contexts where this is a reasonable assumption. Methods for obtaining confidence intervals for the two parameters of the normal distribution, the mean $\mu$ and the standard deviation $\sigma$, are developed. You will see that in order to write down confidence limits for $\mu$, it becomes necessary to make use of a statistical distribution introduced for the first time in this chapter. This is Student’s $t$-distribution, named after W.S. Gosset who published under the pseudonym ‘Student’ at the turn of the century. In this case the calculations are fairly simple and only require reference to statistical tables and a calculator. See Chapter 2, p. 60.

In Section 7.4, it is assumed that samples are large enough for the central limit theorem to be used. This is very commonly the case in practice, and again calculations are based on the normal distribution. Even when (as in Section 7.2) the underlying model is Poisson or binomial, for instance, approximate confidence limits can be calculated using normal tables and a calculator.

In the first four sections of the chapter, it is always assumed that an adequate underlying statistical model (normal, Poisson, binomial, and so on) for the variation observed in a particular context has been identified. Section 7.5 discusses what to do when all you have is a long list of numbers, and not the remotest idea about what might constitute a good model for the underlying variation in the population from which they have been drawn. You will see that, provided your sample is large enough, normal distribution theory can be used to find approximate confidence limits for the underlying population mean.
7.1 Samples of size one

It is not at all common for statistical inferences to be based on just one datum (the singular form of data, 'the thing given'). More often some sort of replication in a statistical experiment is possible, at not too great a cost, and with a consequent enhancement in the precision of the conclusions that may be drawn. However, it will be useful to begin this description of what is involved in setting up a confidence interval by taking a moderately simplified approach in which only one observation has been taken on the random variable of interest.

7.1.1 Some examples

Example 7.1 Accident counts

In an investigation into accident proneness in children, numbers of injuries were counted for 621 children over the eight-year period between the ages of 4 and 11. The early history of each child (aged 4 to 7) was compared with their later history (aged 8 to 11). One child experienced a total of 3 injuries between the ages of 4 and 7. (We shall return to the full data set in Section 7.4.)

For our purposes a Poisson model may be assumed to describe adequately the variation in the number of accidents experienced by children over a four-year period (though, actually, it was an aim of the research exercise to show that a more elaborate model is required in this context). Then the number 3 represents a single observation on the random variable \( N \), say, where \( N \sim \text{Poisson}(\mu) \). The sample mean \( \hat{\mu} = 3 \) is a maximum likelihood estimate of \( \mu \).

This is useful information yielded by the investigation, and this single data point tells us something about the underlying childhood accident rate. It suggests that accidents happen (\( \mu > 0 \)), but not every day, or even every month—perhaps an average of once a year. It would be useful to go further and state, with some confidence, a range of credible values for \( \mu \). Can we devise a procedure that enables us to make a confidence statement along the lines: with 90% confidence, and on the basis of this observation, the value of \( \mu \) lies between 1.5 and 4.8? (Or whatever the confidence limits might turn out to be.)

Example 7.2 Coal-mining disasters

Data were collected on the time intervals in days between disasters in coal mines in Britain from 15 March 1851 to 22 March 1962 inclusive. In this context, an industrial accident is called a 'disaster' if ten or more men were killed. There were altogether 191 such accidents. The data set is a famous one, attracting much interest from analysts, not least because the original published set contained several errors. This kind of delving and reanalysis is a common feature of statistical activity.

The first accident to occur after 15 March 1851 took place on 19 August that year. If data collection had ceased at that point, then the single observation of 157 days would have been collected on a random variable \( T \) which may, for our purposes, be supposed to follow an exponential distribution with unknown


Elements of Statistics

mean \( \mu \). This number provides an estimate of the mean time interval (in days) between consecutive accidents. Can the number also (and, arguably, more usefully) be used to furnish a confidence interval for the average time between consecutive accidents? ■

Example 7.3 Absenteeism

The absences of each of 113 students from a lecture course were recorded over the 24 lectures for which the course ran. There were eleven lectures during the first term and thirteen in the second. Assuming a binomial model for the number of missed lectures in each term, an estimate for the binomial parameter \( p \) (the proportion of lectures missed overall) is provided by the information that during the first term (11 lectures) one student missed four of them. The estimate of \( p \) is \( \hat{p} = 4/11 = 0.364 \). This offers a guide to the underlying value of \( p \), a measure of students’ propensity to miss a lecture. Can the information be developed to provide confidence limits for \( p \)?

An analysis of the extended data set shows that not every student has the same probability for missing a lecture—some students are simply more committed than others—and a binomial model does not provide a good fit. Permitting the absence probability to vary from student to student according to some other statistical distribution could lead to a more realistic, and a better, model. The authors of the paper in which these data were published discuss at some length the fitting of such different models. At the moment, however, the binomial model is as refined a model as we have available to us. For many purposes, it is quite adequate. ■

Example 7.4 Diseased trees

In Chapter 6, Table 6.5 data are given on the lengths of runs of diseased trees in an infected plantation of Douglas firs. The disease was Armillaria root rot, and interest is centred on assessing how infectious the disease was. The geometric distribution turns out to be a good model for the variation in run length. There was one run of 4 infected trees. What can this tell us about the indexing parameter \( p \) for the geometric distribution in this context?

The maximum likelihood estimate of \( p \), based on this single observation, is \( \hat{p} = 1/4 = 0.25 \). It would be useful to develop this into a confidence statement such as: a 95% confidence interval for \( p \), based on the single observation 4, is given by the confidence limits \( p = 0.1 \) to \( p = 0.5 \) (or whatever the limits might be). ■

These four examples all have three features in common. The first is data (3 injuries; 157 days; 4 out of 11; 4 trees). The second is that, in every case, some sort of probability model has been suggested (Poisson, exponential, binomial, geometric). The third feature is that in each case, too, a descriptive model parameter has been identified and related to the data. Let us now pursue the first of these examples further.
Example 7.1 continued

In Example 7.1, a single observation of 3 was recorded on a random variable assumed to follow a Poisson distribution with unknown mean $\mu$. Although the data set is extremely small, it provides an estimate for $\mu$: it is the sample mean, 3. The purpose of this chapter is to answer two questions. The first is: how large could $\mu$ actually be before an observation as low as 3 starts to seem unlikely? This upper limit for $\mu$ is called an upper confidence limit for $\mu$. The second question is similar: how small could $\mu$ actually be before an observation as high as 3 starts to seem unlikely? The answer to this question provides a lower confidence limit for $\mu$. The range of values spanned by the lower and upper confidence limits is called a confidence interval for $\mu$.

To proceed further, we need to decide what in this context is meant by the word ‘unlikely’. For instance, suppose the underlying rate for accidents in early childhood (that is, over the four years from age 4 to age 7) was, in fact, $\mu = 5$. (This is higher than 3, but scarcely renders the observation 3 an incredible one.) In fact, the probability of recording a value as low as that observed (i.e. 3 or lower) is given by the probability $P(X \leq 3)$ where $X \sim \text{Poisson}(5)$. This probability is

$$P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$$

$$= e^{-5} \left( \frac{5^0}{0!} + \frac{5^1}{1!} + \frac{5^2}{2!} + \frac{5^3}{3!} \right)$$

$$= 0.265,$$

which is not particularly small. Extending this argument, if $\mu$ were actually as high as 8, then the probability of observing a count of 3 or lower would be

$$P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$$

$$= e^{-8} \left( \frac{8^0}{0!} + \frac{8^1}{1!} + \frac{8^2}{2!} + \frac{8^3}{3!} \right)$$

$$= 0.042.$$

This is much smaller: there is a probability of only 4% of observing a value as low as 3. If the childhood accident rate $\mu$ were actually as high as 12, then the probability of a child suffering fewer than four accidents would turn out to be just 0.002, which is very small indeed. Inverting that last remark, the observation $x = 3$ suggests that to put the underlying childhood accident rate as high as 12 would be seriously to overestimate it.

What about low values of $\mu$? In this case we need to assess the value of $\mu$ at which an observation as high as 3 (i.e. 3 or higher) starts to look implausible. If we try setting $\mu$ equal to 1, say, then

$$P(X \geq 3) = 1 - P(X \leq 2)$$

$$= 1 - (P(X = 0) + P(X = 1) + P(X = 2))$$

$$= 1 - e^{-1} \left( \frac{1^0}{0!} + \frac{1^1}{1!} + \frac{1^2}{2!} \right)$$

$$= 1 - 0.920$$

$$= 0.080.$$

Of course, it is perfectly possible that the value of $\mu$ is really as high as 12, and an event with a very small probability has happened. This possibility must never be forgotten when constructing confidence intervals.
This is beginning to look a little unlikely. If we try setting \( p \) equal to 0.5, then

\[
P(X \geq 3) = 1 - P(X \leq 2) \\
= 1 - (P(X = 0) + P(X = 1) + P(X = 2)) \\
= 1 - e^{-0.5} \left( \frac{0.5^0}{0!} + \frac{0.5^1}{1!} + \frac{0.5^2}{2!} \right) \\
= 1 - 0.986 \\
= 0.014.
\]

This is even less likely: a value for \( p \) as low as 0.5 is beginning to look implausible, based on the single observation \( x = 3 \). ■

What we have done here is to propose a particular value for \( \mu \) and then to assess, conditional on that value, the chances of observing a count as extreme as the one that actually occurred. If the chances are small, making that observation an unlikely one under the proposed conditions, then we have low confidence in the original proposed value for \( \mu \). A statement about random variables and probability is interpreted in terms of parameters and confidence.

It is a very common practice to express confidence statements in terms of percentages, as in: a 95% confidence interval for \( \mu \), based on the observation \( x = 3 \), is the interval from \ldots \) to \ldots \). There is nothing special about the confidence level of 95%, and you will often see levels of 90% and 99% mentioned. What this means is that an event that might have occurred with probability 0.95 is thought to be 'reasonable', whereas something occurring with probability merely 0.05 will (for the purposes of confident estimation) be considered 'unlikely'.

Now, an event might be considered unlikely because the observed count \( x \) is surprisingly high, or because it is surprisingly low. Let us make these events 'equally' surprising. This suggests that we reserve \( \frac{1}{2} \times 0.05 = 0.025 \) or 2.5% for probability statements about high values of \( x \), and 2.5% for probability statements about low values of \( x \). A 95% confidence interval can be obtained in this case by first solving for \( \mu \) the equation

\[
P(X \leq 3) = e^{-\mu} \left( 1 + \mu + \frac{\mu^2}{2!} + \frac{\mu^3}{3!} \right) = 0.025 \tag{7.1}
\]

(giving the upper confidence limit); and second, solving for \( \mu \) the equation

\[
P(X \geq 3) = 1 - P(X \leq 2) = 1 - e^{-\mu} \left( 1 + \mu + \frac{\mu^2}{2!} \right) = 0.025 \tag{7.2}
\]

giving the lower confidence limit.

The phrase 'solving an equation for \( \mu \)' means rewriting the equation with \( \mu \) as the subject. In other words, the equation has to be rearranged so that \( \mu \) is set equal to an expression that just involves numbers. It is not possible to do this with either (7.1) or (7.2), and other numerical methods need to be followed. As you will discover later in Section 7.2, the first equation has solution \( \mu = 8.8 \), while the second equation has solution \( \mu = 0.62 \). We can complete our calculations with a confidence statement for the unknown Poisson mean \( \mu \), as follows, demonstrating a common notation for lower and upper confidence limits, and the notation that will be used in this course. See Exercise 7.8(a)(ii).
A 95% confidence interval for the Poisson mean $\mu$, based on the single observation 3, is

$$\mu_- = 0.62 \text{ to } \mu_+ = 8.8.$$  

The confidence interval may be written

$$(\mu_-, \mu_+) = (0.62, 8.8).$$

You may have observed that the confidence interval is very wide, with the upper confidence limit nearly fifteen times the lower confidence limit! This is an unavoidable consequence of the sparsity of the data. We shall see that larger samples typically reduce the width of confidence intervals very usefully.

The approach may be summarized as follows. A single observation $x$ is collected on the random variable $X$, where $X$ follows a specified probability model with unknown parameter $\theta$. Then a 95% confidence interval for $\theta$ is provided by solving separately for $\theta$ the equations

$$P(X \leq x) = 0.025 \text{ and } P(X \geq x) = 0.025.$$  

It has already been remarked that there is nothing special about the confidence level of 95%, although it is a common one to choose. Depending on the purpose for which confidence statements are formulated, selected levels might be as low as 90% or as high as 99% or 99.9%. It is common to write the level generally as $100(1 - \alpha)%$, and to solve for $\theta$ the two equations

$$P(X \leq x) = \frac{1}{2}\alpha \text{ and } P(X \geq x) = \frac{1}{2}\alpha.$$  

In general, the arithmetic of such computations is difficult, and numerical techniques beyond the scope of this course have to be adopted. Alternatively, many statistical computer packages make such techniques unnecessary, returning confidence limits at the press of a key. The exercises of Section 7.2 assume that you have access to such a computer package.

### Confidence intervals

Suppose that a single observation $x$ is collected on a random variable $X$, following a specified probability distribution with unknown parameter $\theta$. Then a confidence interval $(\theta_-, \theta_+)$ for $\theta$ with level $100(1 - \alpha)%$ is provided by separately solving for $\theta$ the two equations

$$P(X \leq x) = \frac{1}{2}\alpha \text{ and } P(X \geq x) = \frac{1}{2}\alpha.$$  

In some cases the arithmetic is straightforward, as in the following example.

#### Example 7.2 continued

In the example on coal-mining disasters, the single number 157 (days) was obtained as an observation on an exponentially distributed random variable $T$ with mean $\mu$. Suppose that what is now required is a 90% confidence interval for $\mu$, based on the single observation $t = 157$. The procedure is as follows.

The required confidence level is $100(1 - \alpha)% = 90\%$, so $\frac{1}{2}\alpha = 0.05$. For an exponentially distributed random variable $T$ with mean $\mu$, the c.d.f. is given here, both the upper and lower confidence limits have been given to two significant figures (rather than, say, two decimal places).
by
\[ P(T \leq t) = 1 - e^{-t/\mu}, \quad t \geq 0. \]
The equation \( P(T \leq t) = 0.05 \), which yields an upper confidence level for \( \mu \) in the exponential case, may be solved as follows.
\[ P(T \leq t) = 1 - e^{-t/\mu} = 0.05 \]
gives
\[ e^{-t/\mu} = 0.95 \]
so
\[ \frac{t}{\mu} = -\log(0.95) \]
or
\[ \mu = \frac{t}{-\log(0.95)} = \frac{157}{0.051293} = 3060.8 \]
or, say,
\[ \mu_+ = 3060 \text{ days.} \]

**Exercise 7.1**

Calculate the corresponding lower confidence limit \( \mu_- \) for \( \mu \), the average number of days between disasters, assuming an exponential model and based on the single observation 157 days. Write out explicitly a confidence statement about the value of \( \mu \).

Notice that in Exercise 7.1 it is again evident that calculation of confidence limits requires three things: data, a model and a statement of the parameter of interest.

Another feature you might have noticed is the extreme width of the confidence interval that results! (This was true for the Poisson mean as well.) The time interval observed was about five months; the confidence limits for the mean time between disasters are from 'rather less than two months' to 'more than eight years'. This confidence interval is not very informative, due partly to the highly skewed shape of the exponential distribution, but, again, mainly to the dearth of data.

Here is another example where the arithmetic is straightforward.

**Example 7.5 A crooked die**

A gambler, who believes that a die has been loaded in such a way that rolling a 6 is less likely than it ought to be (i.e. \( p < 1/6 \)), discovers in a single experiment that it takes 13 rolls to obtain a 6. If the die is fair, then the number of throws necessary to achieve the first 6 is a random variable \( N \) following a geometric distribution with parameter \( p = 1/6 \). The expected value of \( N \) is \( \mu_N = 1/p = 6 \). Certainly, an observation as high as 13 suggests that \( p \) is smaller than 1/6; indeed, the maximum likelihood estimate of \( p \) is
\[ \hat{p} = \frac{1}{13} = 0.077. \]
But it is quite possible for an observation as high as 13 to be observed even if the die is fair. Here, in Table 7.1 are some simulated observations on the random variable \( N \sim G(p) \), where \( p = \frac{1}{6} \).

In none of these twenty simulations was a count as high as 13 observed; but there were two 11s and one 12. For a random variable \( N \) following a geometric distribution with parameter \( p \), the c.d.f. of \( N \) is given by

\[
P(N \leq n) = 1 - (1 - p)^n, \quad n = 1, 2, 3, \ldots.
\]

If \( p = \frac{1}{6} \), an observation as extreme as 13 (i.e. 13 or more) has probability

\[
P(N \geq 13) = 1 - P(N \leq 12) \\
= 1 - \left(1 - (1 - \frac{1}{6})^{12}\right) \\
= \left(\frac{5}{6}\right)^{12} \\
= 0.112,
\]

so an observation as large as 13 is not extraordinary. What we could do to try to clarify matters for the gambler is to calculate confidence limits for \( p \) based on the single datum 13. Let us set the confidence level required at 95%. Then \( 100(1 - \alpha)\% = 95\% \) so \( \frac{1}{2} \alpha = 0.025 \). First, we solve for \( p \) the equation

\[
P(N \leq 13) = 1 - (1 - p)^{13} = 0.025.
\]

The solution is

\[
p = 1 - (1 - 0.025)^{1/13} = 1 - 0.998 = 0.002,
\]

so \( p_- = 0.002 \).

(Notice that in this case, solving the equation \( P(N \leq n) = \frac{1}{2} \alpha \) has yielded a lower confidence limit for the parameter: for here, it is low values of \( p \) that render low values of \( N \) unlikely.)

### Exercise 7.2

Assuming a geometric model, find the corresponding upper confidence limit \( p_+ \) for \( p \), and complete your confidence statement for \( p \), based on the single observation \( n = 13 \).

### Exercise 7.3

Find a 90% confidence interval for \( p \), and compare it with your 95% confidence interval.

So the 95% confidence interval for \( p \), based on these data and assuming an underlying geometric model, is given by

\[
(p_-, p_+) = (0.002, 0.265),
\]

and the 90% confidence interval is

\[
(p_-, p_+) = (0.004, 0.221).
\]

Notice particularly that the number \( p = 1/6 = 0.167 \) is included both in the 90% and 95% confidence intervals, which are extremely wide: the case for
'collecting more data' is seriously pressing. There seems no particular reason to suppose that the die is unfairly loaded away from 6s, even though the experiment in Example 7.5 took 13 rolls of the die to achieve a 6. The idea of performing statistical experiments to explore propositions is a very important one, and one that is pursued in Chapter 8.

**Exercise 7.4**

In Example 7.4, one run of diseased trees was of length 4. Assuming a geometric model for the distribution of run length, use this observation to construct a 99% confidence interval for the mean length $\mu$ of runs of diseased trees.

**Note** Here, you can establish confidence limits for $p$ and then simply make use of the relationship for a geometric random variable that $\mu = 1/p$ to rewrite the upper and lower confidence limits. Or you could reparameterize the geometric distribution in terms of $\mu$ rather than $p$ by saying

$$P(N \leq n) = 1 - (1 - p)^n = 1 - \left(1 - \frac{1}{\mu}\right)^n,$$

and then work directly in terms of the parameter $\mu$.

**Example 7.3 continued**

The only one of the four examples with which this section was introduced and which has not yet been followed through is Example 7.3, in which 4 absences were observed in a total of 11 lectures. The maximum likelihood (and common-sense) estimate for $p$, the underlying probability of absence, is $\hat{p} = 4/11 = 0.364$. Again, it would be useful to develop this estimate into a confidence interval for $p$. Assuming independence from lecture to lecture (which is probably rather a strong assumption in this context) then we have a single observation $x = 4$ on a binomially distributed random variable $X \sim B(11, p)$. We need therefore to solve the two equations

$$P(X \leq 4) = \frac{1}{2}\alpha$$

and

$$P(X \geq 4) = \frac{1}{2}\alpha$$

for $p$. The first reduces to

$$P(X \leq 4) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$$

$$= \binom{11}{0} p^0 q^{11} + \binom{11}{1} p^1 q^{10} + \binom{11}{2} p^2 q^9 + \binom{11}{3} p^3 q^8 + \binom{11}{4} p^4 q^7$$

$$= (1 - p)^{11} + 11p(1 - p)^{10} + 55p^2(1 - p)^9 + 165p^3(1 - p)^8 + 330p^4(1 - p)^7$$

$$= \frac{1}{2}\alpha.$$

The algebra here is not at all convenient (as it has been in the case of exponential and geometric probability calculations). The equation reduces to

$$(1 - p)^7(210p^4 + 84p^3 + 28p^2 + 7p + 1) = \frac{1}{2}\alpha,$$

which is not soluble by usual methods. Actually, the 90% upper confidence limit ($\frac{1}{2}\alpha = 0.05$) is given by $p_r = 0.6502$. This kind of calculation can be
executed with a single command if you have a computer with the appropriate software installed. The second equation \( P(X \geq 4) = \frac{1}{2} \alpha \) reduces to

\[
1 - (1 - p)^8 (120p^3 + 36p^2 + 8p + 1) = \frac{1}{2} \alpha.
\]

This equation has the solution (when \( \frac{1}{2} \alpha = 0.05 \)) \( p_\alpha = 0.1351 \). So we might conclude: a 90% confidence interval for \( p \), based on observing 4 successes in 11 trials and assuming independence from trial to trial, is given by \((p_\alpha, p_+) = (0.1351, 0.6502) \).

**7.1.2 Interpreting a confidence interval**

You have seen that a requirement to construct a confidence interval for an unknown parameter \( \theta \) is the collection of data. In different random samples, the data will vary. It follows (rather as in the case of point estimation) that the resulting confidence limits \( \theta_- \) and \( \theta_+ \) are themselves random variables. The resulting confidence interval \((\theta_-, \theta_+)\) is called a random interval. What are its relevant properties?

This question will be explored in one very simple case. The principle is the same in other sampling contexts, but the algebra can become rather involved.

Suppose that a single observation \( X \) has been collected from a population assumed to follow an exponential distribution with mean \( \mu \). The 2.5% point of this distribution is the solution \( q_{0.025} \) of the equation \( 1 - e^{-q/\mu} = 0.025 \), or \( q_{0.025} = -\mu \log(0.975) = 0.025\mu \). Similarly, the 97.5% point is \( q_{0.975} = 3.69\mu \). So we can write

\[
P(0.025\mu \leq X \leq 3.69\mu) = 0.95.
\]

Then the double inequality on the left-hand side can be rewritten with \( \mu \) as the subject:

\[
P \left( \frac{X}{3.69} \leq \mu \leq \frac{X}{0.025} \right) = P(0.27X \leq \mu \leq 39.5X) = 0.95.
\]

This is a statement about the random interval \( 0.27X \) to \( 39.5X \): it is a very wide interval, stretching from about one-quarter of the single observed data point to forty times it. The point is that with probability 0.95, the random interval contains the unknown number \( \mu \).

That is to say: if the statistical experiment were repeated many times, and if the random interval

\[
(\mu_-, \mu_+) = (0.27X, 39.5X)
\]

were computed each time, then approximately 95% of these intervals would contain the unknown number \( \mu \). It is a common error to complete the statistical investigation with a 'confidence' statement such as: with probability 0.95, the value of \( \mu \) is between \( \mu_- \) and \( \mu_+ \). But the number \( \mu \) is not a random variable. It is important to remember that a confidence statement follows from a probability statement about a random interval, which might, or might not, contain the constant \( \mu \).
### 7.1.3 Some further exercises

**Exercise 7.5**

In Chapter 2, Example 2.8, a triangular model $T \sim \text{Triangular}(20)$ was suggested for the waiting time (in seconds) between consecutive vehicles using a particular grade of road.

(a) Find in terms of $c_w$ and $t$ a defining formula for the $100(1 - \alpha)\%$ confidence interval for the parameter $\theta$ of a triangular density, based on the single observation $t$.

(b) Use your formula to give a 95% confidence interval for the parameter $\theta$ of a triangular distribution, based on the single observation $t = 5$.

**Exercise 7.6**

Firefly racing dinghies carry sail numbers consisting of a capital letter F followed by one or more digits. The dinghies are numbered consecutively as they are manufactured, starting at F1.

The total number of Firefly dinghies manufactured is not known. However, one dinghy has been sighted. It bore the sail number F3433. On the basis of this one observation, and stating your model for the population from which this one observation was drawn, calculate a 90% confidence interval for the total number of Firefly dinghies manufactured to date.

### 7.2 Small samples

In Section 7.1 only very small data sets (samples of size one) were considered. In all cases where you had to calculate a confidence interval for some unknown model parameter, one recurring feature was the extreme width of the intervals found. In this section and in Section 7.4 you will see the beneficial consequences of using larger samples.

As in Section 7.1, the derivation of a confidence interval for a population characteristic requires data, and specification of a model for the inherent variation. The characteristic of interest will be expressible in terms of the model parameter.

In most of the cases that we shall look at in this section, the underlying statistical theory is based on the sample total; for the sort of models used in this course much is known about the distribution of sums of random variables. In Section 7.4 the samples will be assumed to be sufficiently large for the central limit theorem to be implemented: the sample total may be assumed to be approximately normally distributed. However, in this section, the samples are small and the exact distribution of the sample total is used. Brief explanations are given of the principles on which interval estimation in this section is based, but the details are attended to by the computer.

However, the sample total is not always the appropriate statistic to use, and at the end of this section you will see an example where this is the case.

You have already seen that even when the sample properties are well known and easy to write down, the arithmetic involved in calculating confidence
limits is not necessarily straightforward. We have so far been unable directly to
deduce confidence limits for a binomial probability $p$ or a Poisson mean $\mu$,
for instance, although you have seen examples of them.

The exercises in this section require the use of a computer to cope with the
arithmetic. The problems are stated: you need to explore and use the facilities
of your computer in order to solve them.

### 7.2.1 Confidence intervals for the Bernoulli parameter

A researcher interested in the proportion $p$ of members of a population who
possess a particular attribute is not likely to base an assessment of the value
of $p$ on a single observation $X$, taking the value 0 (attribute not possessed) or
1 (attribute possessed). It is more likely that a larger sample $X_1, X_2, \ldots, X_n$
($n > 1$) will be collected. Assuming independence between sample responses,
then the sample total $X_1 + X_2 + \cdots + X_n$ follows a binomial distribution
$B(n, p)$ and an estimate of $p$ is provided by

$$\hat{p} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

We have already seen in principle how to extract a $100(1 - \alpha)$% confidence
interval for $p$, based on a statistical experiment in which $x$ successes were observed in a total of $n$ trials.

**Exercise 7.7**

Use your computer to calculate confidence intervals for the Bernoulli parameter $p$ in each of the following situations.

(a) A total of 4 successes is observed in 11 trials (so $\hat{p} = 4/11 = 0.364$). Find
(i) a 90% confidence interval for $p$ and (ii) a 95% confidence interval for
$p$ based on these data.

(b) In a similar follow-up experiment, the sample size is doubled to 22; and,
as it happens, a total of 8 successes is observed. So $\hat{p} = 0.364$ as before.
Find (i) a 90% confidence interval for $p$ and (ii) a 95% confidence interval
for $p$ based on these data. Can you explain why using a larger sample
provides narrower confidence intervals?

(c) Find a 99% confidence interval for a Bernoulli parameter $p$ based on
observing 4 successes in 5 trials.

(d) Simulation Use your computer to generate 10 observations on the bi-
nomial distribution $B(20, 0.3)$. (You will obtain a sequence of success
counts—3 out of 20, 8 out of 20, 2 out of 20, \ldots, and so on.) For each of
these 10 experimental results, calculate the corresponding 90% confidence
interval for $p$. How many of your intervals contain the value 0.3? How
many would you have expected to contain the value 0.3?

### 7.2.2 Confidence intervals for a Poisson mean

You have seen the sort of calculation involved in obtaining confidence limits
for a Poisson parameter $\mu$ when a single data point is available. What
happens, when more than one data point is sampled, is quite straightfor-
ward: the calculations are almost the same. If \( n \) independent observations \( X_1, X_2, \ldots, X_n \) are taken from a Poisson distribution with mean \( \mu \), then

their sum \( T = X_1 + X_2 + \cdots + X_n \) follows a Poisson distribution with mean \( \mu_T = n\mu \). A confidence interval is found for \( \mu_T \) based on the single observation \( t \). Then the confidence limits for \( \mu_T \) can be converted to confidence limits for \( \mu \) simply by dividing by \( n \), the sample size.

### Exercise 7.8

Use your computer to calculate confidence intervals for the Poisson parameter \( \mu \) in each of the following situations.

(a) A child is observed for one year, during which period he suffers 3 minor accidents. Use this information to calculate (i) a 90% confidence interval and (ii) a 95% confidence interval for the underlying annual accident rate for boys of his age.

(b) Seven boys, all of the same age, are observed for a year. They suffer 4, 4, 3, 0, 5, 3, 2 minor accidents respectively. Use this information (i) to estimate the underlying annual accident rate \( \mu \) for boys of their age and (ii) to calculate 90% and 95% confidence intervals for \( \mu \).

(c) Six girls, all of the same age, are observed for a year. Between them they suffer a total of 20 minor accidents. Use this information (i) to estimate the underlying annual accident rate \( \mu \) for girls of their age and (ii) to calculate 90% and 95% confidence intervals for \( \mu \).

Notice that in part (b) of this exercise all the information was given; in part (c) only summary data were provided.

### 7.2.3 Confidence intervals for an exponential mean

If \( n \) independent observations are taken from an exponential distribution with mean \( \mu \), then their sum follows a probability distribution which is a member of the two-parameter **gamma** family. In this case the values of the two parameters are \( n \) and \( \mu \), but the details are, in this context, unimportant. Enough is known about relevant properties of the gamma distribution to enable computer routines to be written for the evaluation of confidence limits for exponential parameters based on more than one observation.

### Exercise 7.9

(a) The data in Chapter 4, Table 4.7 give 62 time intervals in days between consecutive serious earthquakes world-wide. Use these data to construct a 90% confidence interval for the mean time interval between earthquakes, stating any probability model you assume.

(b) The data listed in Chapter 2, Table 2.11 are the time intervals (in seconds) between successive vehicles using the Kwinana Freeway one morning in Perth, Western Australia.

(i) Use these data to estimate the rate (vehicles per minute) at which vehicles pass the observation point. (ii) Calculate a 90% confidence interval for the rate.
The last part of this exercise involves a computer simulation.

(c) Generate 20 observations from the exponential distribution $M(1)$, and use your data to estimate the (usually unknown) population mean; then use the data to calculate a 90% confidence interval for the population mean. Repeat the process 10 times. In each case, examine whether the resulting confidence interval contains the number 1.

### 7.2.4 Confidence intervals for a geometric parameter

If $n$ independent observations are taken from a geometric distribution with parameter $p$, then their sum follows a probability distribution which is a member of the two-parameter **negative binomial** family. The values of the two parameters are $n$ and $p$, but again the details in this context are unimportant. Enough is known about relevant properties of the negative binomial distribution to enable the development of computer routines for the evaluation of confidence limits for geometric parameters based on more than one observation.

#### Exercise 7.10

In this exercise you will need to use your computer to count the number of rolls $N$ necessary to achieve a 6, when a loaded die with $P(6) = 1/10$ is rolled.

(a) Obtain 10 observations on $N$, and use these observations to obtain a 90% confidence interval for the (usually unknown) probability $p = P(6)$. Does your interval contain the value $p = 1/10$? Does it contain the value $p = 1/6$?

(b) Now obtain 100 observations on $N$, and use them to obtain a 90% confidence interval for $p$. How does the width of this interval compare with the width of the interval you obtained in part (a)? Does your new interval contain the value $p = 1/10$? Does it contain the value $p = 1/6$?

### 7.2.5 Postscript

So far in this section we have seen the useful consequences of drawing a larger sample of data in order to calculate confidence intervals for unknown model parameters: the larger the sample, the narrower (that is, the more precise) the resulting interval tends to be.

Your computer will have (or should have) insulated you from the algebraic detail of the computations involved, but for all four of the standard cases considered so far, inference was based on the sample total. We have not considered samples from the triangular or Pareto distributions, or from the uniform distribution, all of which you have met in the course so far. It is sometimes far from easy to obtain the distribution of the sum $X_1 + X_2 + \cdots + X_n$ of a random sample from some stated distribution. In such a case the only thing to do (if you cannot calculate exact results numerically) is to make sure that the sample drawn is so large that the sample total may be assumed to be approximately normally distributed (by the central limit theorem). We shall look at this in Section 7.4.
To round off this section, we shall look at a situation where the sample total is not an appropriate statistic to use when calculating a confidence interval, but where an exact computation is nevertheless possible.

The uniform distribution is rather unusual. Recall the Firefly example in Exercise 7.6. If, say, five Firefly dinghies had been sighted (rather than just one), then would that have helped us further? Suppose the numbers sighted had been 3433 (as before) and then 1326, 378, 1826, 1314. It might seem as though the numbers 1326, 378, 1826, and 1314 really provide no further information at all about the total number of Firefly dinghies manufactured altogether: that number is evidently at least 3433, as was apparent from the first sighting. In fact, a useful inference can be based not on the sample total but on the sample maximum. The maximum $X_{\text{max}}$ of a random sample of size $n$ from a discrete uniform distribution $U(1,2,\ldots,\theta)$ has c.d.f.

$$P(X_{\text{max}} \leq x) = \left(\frac{x}{\theta}\right)^n, \quad x = 1,2,\ldots,\theta.$$ 

The corresponding 100$(1 - \alpha)$% confidence interval for $\theta$ based on $n$ observations with maximum value $x$ is found by solving the two equations

$$P(X_{\text{max}} \leq x) = \left(\frac{x}{\theta}\right)^n = \frac{1}{2} \alpha$$

and

$$P(X_{\text{max}} \geq x) = 1 - P(X_{\text{max}} \leq x - 1) = 1 - \left(\frac{x-1}{\theta}\right)^n = \frac{1}{2} \alpha.$$ 

So, for instance, the 90% confidence interval for $\theta$ based on the five observations with maximum 3433 is found by solving the two equations

$$P(X_{\text{max}} \leq 3433) = \left(\frac{3433}{\theta}\right)^5 = 0.05,$$

which has solution $\theta_+ = 3433/0.05^{1/5} = 6250$; and

$$P(X_{\text{max}} \geq 3433) = 1 - P(X_{\text{max}} \leq 3432) = 1 - \left(\frac{3432}{\theta}\right)^5 = 0.05,$$

with solution $\theta_- = 3432/0.95^{1/5} = 3467$. The corresponding confidence interval for $\theta$ is given by $(\theta_-, \theta_+) = (3467, 6250)$. If you compare this with your result in Exercise 7.6 you can see that, again, a larger sample has resulted in a more useful confidence interval (because it is narrower).

### 7.3 Confidence intervals for the parameters of a normal distribution

In the previous sections, we did not consider the fundamental problem of constructing confidence intervals for the two parameters of a normal population, based on a random sample drawn from that population. This should now be quite straightforward: all that is required is to write down probability statements about statistics derived from the random sample (such as the sample mean and sample standard deviation) and invert those statements so that the unknown parameters $\mu$ and $\sigma$ become the subjects of confidence statements.
### 7.3.1 Confidence intervals for the normal mean

Here is a typical example of the sort of statistical experiment which arises, where a normal model may reasonably be assumed, but where nothing is known either about the indexing mean $\mu$ or about the standard deviation $\sigma$.

#### Example 7.6 A mechanical kitchen timer

A kitchen timer is a small alarm clock that, by turning a dial, can be set to ring after any length of time between one minute and an hour. It is useful as a reminder to somebody working in a kitchen that some critical stage has been reached. The usefulness of such timers is not restricted to the kitchen, of course.

An enthusiastic cook was interested in the accuracy of his own kitchen timer, and on ten different occasions set it to ring after a five-minute delay (300 seconds). The ten different time intervals recorded on a stop-watch are shown in Table 7.2.

Assuming that the stop-watch itself was an accurate measuring instrument, then the only variability from the 300 seconds' delay intended in the times recorded is due to difficulties in actually setting the time (that is, positioning the dial) and to mechanical malfunction in the operation of the timer. Assuming a normal model $N(\mu, \sigma^2)$ for the variation in the actual times recorded when a time of five minutes is set, then the data yield the parameter estimates

$\bar{x} = 294.81, \quad s^2 = 3.1232$.

That is, an estimate of the normal mean $\mu$ is the sample mean $\hat{\mu} = \bar{x} = 294.81$ seconds (about $4\text{m} 55\text{s}$—five seconds short of the five-minute interval set); an estimate of the normal standard deviation $\sigma$ is the sample standard deviation $s = \sqrt{3.1232} = 1.77$ seconds.

Can we say more about the average time delay other than that it seems to be about $4\text{m} 55\text{s}$? Perhaps an actual mean of five minutes (the delay intended) is plausible ... but, given these data, how confident could one be in putting this proposition forward?

If $\bar{X}$ is the mean of a random sample of size $n$ from a normal distribution with mean $\mu$ and standard deviation $\sigma$, then the distribution of $\bar{X}$ is given by

$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$;

on standardizing, this is equivalent to

$Z = \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right) \sim N(0, 1)$.

This result enables us to write down probability statements of the general form

$P\left(-z \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z\right) = 1 - \alpha$. 

---

Data provided by B.J.R. Bailey, University of Southampton.

**Table 7.2** Ten time delays (seconds)

<table>
<thead>
<tr>
<th>Time Delay (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>293.7</td>
</tr>
<tr>
<td>296.2</td>
</tr>
<tr>
<td>296.4</td>
</tr>
<tr>
<td>294.0</td>
</tr>
<tr>
<td>297.3</td>
</tr>
<tr>
<td>293.7</td>
</tr>
<tr>
<td>294.3</td>
</tr>
<tr>
<td>291.3</td>
</tr>
<tr>
<td>295.1</td>
</tr>
<tr>
<td>296.1</td>
</tr>
</tbody>
</table>

---

281
Based on what we know about normal probabilities, we might write, say,

\[ P \left( -1.96 \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq 1.96 \right) = 0.95. \]

This follows since 95% of observations on a normal population are within 1.96 standard deviations of the mean. If the population mean \( \mu \) was the only unknown quantity in this probability statement we could rewrite the double inequality with \( \mu \) as the subject, as follows:

\[
0.95 = P \left( -1.96 \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq 1.96 \frac{\sigma}{\sqrt{n}} \right)
= P \left( -1.96 \frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq 1.96 \frac{\sigma}{\sqrt{n}} + \mu \right)
= P \left( \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right). 
\]

Thus we obtain a random interval, centred on the sample mean \( \bar{X} \), which with probability 0.95 contains the unknown population mean \( \mu \). So, apparently, a 95% confidence interval for \( \mu \) based on a sample \( x_1, x_2, \ldots, x_n \) with sample mean \( \bar{x} \), is given by

\[
(\mu_-, \mu_+) = \left( \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \right).
\]

This result is useful if the normal standard deviation \( \sigma \) is known. Usually, however, in any context where the mean is unknown then the standard deviation will be unknown as well. So this approach cannot be used for making inferences about the unknown parameter \( \mu \), because of the presence of the unknown term \( \sigma \). What is to be done? The obvious solution is to replace \( \sigma \) by its estimator \( S \), the sample standard deviation. Remember that \( S \) is the square root of the sample variance

\[ S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \]

and \( E(S^2) = \sigma^2 \), so the substitution seems quite sensible.

Let us now define a new random variable \( T \) (as at (7.3)) by

\[ T = \frac{\bar{X} - \mu}{S/\sqrt{n}}. \quad (7.4) \]

In this expression the right-hand side has the property that the unknown parameter \( \sigma \) does not feature. We can then go on to make probability statements of the kind

\[ P \left( -t \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t \right) = 1 - \alpha \]

(exactly as before), but first we have to find the appropriate values of \( t \). That is, before we can use the random variable \( T \) to make inferences about \( \mu \), we
need to know its properties, and in particular its probability distribution.
Since we have obtained $T$ from $Z$ by replacing a constant, $\sigma$, by a random
variable, $S$, you will not be surprised to be told that $T$ is not normally dis-
tributed.

In fact there is not just one distribution for $T$, but a whole family of distri-
butions, one for each value of the sample size $n$, ($n = 2,3,\ldots$).

The distribution is often known as **Student’s $t$-distribution** since it was first derived, in 1908, by W.S. Gosset who wrote under the pseudonym of ‘Student’. The family is indexed by means of a parameter called the degrees
of freedom (just as in the case of a $\chi^2$ variate—see Chapter 6). Thus we write
$T \sim t(\nu)$ to denote that the random variable $T$ has a $t$-distribution with \( \nu \) degrees of freedom.

Like the standard normal variate $Z$ each member of this family of distribu-
tions, for $\nu = 1,2,3,\ldots$, is symmetric about 0: the numerator of $T$, the difference $\overline{X} - \mu$, is as likely to be negative as it is to be positive. But in view of its dependence on two sample statistics (the sample standard deviation $S$
as well as the sample mean $\overline{X}$) the random variable $T$ is, in a sense, more
variable than $Z$, and its probability density function has ‘longer tails’ than
that of $Z$. All this is clearly seen in Figure 7.1 which shows the probability
density function of $Z$ together with those of $t(1)$, $t(3)$ and $t(7)$—that is, the
t-distributions with 1, 3 and 7 degrees of freedom.

![Figure 7.1](image)

*Figure 7.1* The densities of $t(1), t(3), t(7)$ and $Z \sim N(0,1)$

The larger the value of $\nu$, the closer is the distribution of $t(\nu)$ to a standard normal distribution. This is illustrated in the comparative sketches of the corresponding probability density functions in Figure 7.1. This makes sense:
for a large sample size, $S$ should be an accurate estimator for $\sigma$, and the distribution of the random variable $T$ will be close to the distribution of the standard normal variate $Z$.

In order to make inferences about the unknown parameter $\mu$ using Student’s $t$-distribution, it is necessary to obtain critical values from statistical tables.

A table of critical values of $t(\nu)$ for different values of $\nu$ is given as Table A5 in the Appendix. As in the case of the normal distribution, a sketch always helps. Here, in Figure 7.2, are some examples showing different critical values for Student’s $t$-distribution, for different degrees of freedom $\nu$. You should check that you can find these critical values using the table.
Figure 7.2 Critical values of $t(\nu)$

For example, in Figure 7.2(a), the size of the right-hand shaded area is $\frac{1}{2}(1 - 0.95) = 0.025$: the corresponding critical value is the 97.5% point ($q_{0.975}$) of $t(9)$. Looking along the ninth row of the table (corresponding to $\nu = 9$) and under the column corresponding to the probability $P(T \leq t) = 0.975$, the value of $t$ is 2.262.

You should check that you can use the table to answer the questions in the exercise that follows, and confirm your findings by typing the appropriate command at your computer.

Exercise 7.11

(a) If $\nu = 29$, determine $t$ such that $P(T \leq t) = 0.95$.
(b) If $\nu = 30$, determine $t$ such that $P(T \geq t) = 0.05$.
(c) If $\nu = 5$, determine $t$ such that $P(T \leq t) = 0.01$.
(d) If $\nu = 9$, determine $t$ such that $P(|T| \leq t) = 0.95$.

You have just used tables of the $t$-distribution and your computer to determine the value of stated percentage points; that is, the value of the critical point $t$ which (for given values of $\nu$ and $\alpha$) will permit probability statements of the form

$$P(-t \leq T \leq t) = 1 - \alpha.$$
where \( T \sim t(\nu) \). Now, suppose in a sampling context that \( \bar{X} \) is the mean of a random sample from a normal distribution with mean \( \mu \), and that \( S \) is the sample standard deviation. In this context, the random variable \( T \) defined at (7.4) has Student's \( t \)-distribution with degrees of freedom \( \nu = n - 1 \).


In a random sample of size \( n \) from a normal distribution with mean \( \mu \), the random variable

\[
T = \frac{\bar{X} - \mu}{S/\sqrt{n}}
\]

(where \( \bar{X} \) is the sample mean and \( S \) is the sample standard deviation) follows **Student's \( t \)-distribution** (or simply a \( t \)-distribution) with \( n - 1 \) degrees of freedom. This is written

\[
T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n - 1).
\]

An important consequence of this result is that it is possible to make probability statements of the form

\[
P \left( -t \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t \right) = 1 - \alpha,
\]

where \( t \) is the \( 100(1 - \frac{1}{2}\alpha)\% \) point of \( t(n - 1) \). This probability statement may be rewritten

\[
P \left( \bar{X} - \frac{tS}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{tS}{\sqrt{n}} \right) = 1 - \alpha,
\]

which is a probability statement about a random interval, centred on the sample mean \( \bar{X} \), which with probability \( 1 - \alpha \) contains the unknown population mean \( \mu \). The corresponding confidence interval for \( \mu \), based on a random sample \( x_1, x_2, \ldots, x_n \), is

\[
(\mu_-, \mu_+) = \left( \bar{x} - \frac{ts}{\sqrt{n}}, \bar{x} + \frac{ts}{\sqrt{n}} \right).
\]

Here, \( \bar{x} \) is the sample mean, \( s \) is the sample standard deviation, \( n \) is the sample size and \( t \) is a critical value obtained from tables of Student's \( t \)-distribution (or from your computer).

**Example 7.6 continued**

If it is assumed that the time until the kitchen timer alarm bell sounds is normally distributed, but with unknown mean and variance, then we are now in a position to construct a confidence interval for the unknown mean waiting time \( \mu \). We know that the sample statistics are \( \bar{x} = 294.81 \) and \( s = \sqrt{3.1232} = 1.77 \). For a 90\% confidence interval, say, with \( \nu = n - 1 = 9 \)
degrees of freedom, the critical \( t \)-value is 1.833 (using \( \frac{1}{2} \alpha = 0.05 \)). Hence the confidence interval we obtain for \( \mu \), based on these data, is

\[
(\mu_-, \mu_+) = \left( \bar{x} - \frac{ts}{\sqrt{n}}, \bar{x} + \frac{ts}{\sqrt{n}} \right)
\]

\[
= \left( 294.81 - \frac{1.833 \times 1.77}{\sqrt{10}}, 294.81 + \frac{1.833 \times 1.77}{\sqrt{10}} \right)
\]

\[
= (294.81 - 1.03, 294.81 + 1.03)
\]

\[
= (293.8, 295.8).
\]

Notice that the number 300 (seconds, a five-minute delay) is not contained within this confidence interval! (Also, the confidence interval is usefully narrow.) So there is some statistical evidence here that the timer is consistently going off early. We shall return to this very important point in Chapter 8. ■

Exercise 7.12

In 1928 the London and North-Eastern Railway ran the locomotive Lemberg with an experimental boiler pressure of 220 lb, in five trial runs. Several random variables were measured, one being the coal consumption, in pounds per draw-bar horse-power hour; the resulting observations were

\[
3.27 \quad 3.17 \quad 3.24 \quad 2.92 \quad 2.99.
\]

Regarding these data as a random sample from a normal distribution, construct a 95% confidence interval for the population mean \( \mu \) (that is, the mean coal consumption at a boiler pressure of 220 lb).

Exercise 7.13

The following data (see Table 7.3) are from ‘Student’s’ 1908 paper. Several illustrative data sets were used in the paper; this one is taken from a table by A.R. Cushny and A.R. Peebles in the Journal of Physiology (1904) showing the effects of the optical isomers of hyoscyamine hydrobromide in producing sleep.

The sleep of ten patients was measured without hypnotic and after treatment (a) with D-hyoscyamine hydrobromide and (b) with L-hyoscyamine hydrobromide. The average number of hours’ sleep gained by the use of the drug was tabulated. Here, D- denotes dextro and L- denotes laevo, a classification system for stereoisomers.

The table is not exactly as printed in Biometrika. It is a pity that in this fundamental paper there should have been a typographical error: this has been corrected in Table 7.3 (and it was not difficult to isolate). A minus sign indicates a net sleep loss.

For the purposes of this exercise, only consider the ten numbers in the column headed L-D. This summarizes any differences between the two treatments. (Looking at the values it seems as though treatment with L-hyoscyamine hydrobromide was consistently the more effective.)
(a) Calculate the sample mean $\bar{x}$ and the sample standard deviation $s$ for these data, and hence construct a 95% confidence interval for the unknown mean difference $\mu$, assuming the differences to be normally distributed.

(b) Say whether the number 0 is contained in the confidence interval you found in part (a). Can you deduce anything from your study?

In general, for larger samples, the calculation of the statistics $\bar{x}$ and $s$ may be inconvenient. Not only can a computer perform this arithmetic task, but it can then go on to 'look up critical percentage points' against the appropriate $t$-distribution and hence calculate confidence intervals. You do not have to do anything at all, other than key in the appropriate command. Your computer probably carries out this exercise quite easily.

**Exercise 7.14**

Confirm your answers to Exercises 7.12 and 7.13 by obtaining the confidence intervals sought directly from your computer.

### 7.3.2 Confidence intervals for the normal variance

The normal distribution is indexed by two parameters: the second is its variance, denoted $\sigma^2$. In contexts where this quantity is unknown, it may often be useful to draw a random sample of observations in order to obtain a confidence interval. Let us start with an example.

**Example 7.7 Breaking strengths**

In the production of synthetic fibres, it is important that the fibres produced are consistent in quality. One aspect of this is that the tensile strength of the fibres should not vary too much. A sample of eight pieces of fibre produced by a new process is taken, and the tensile strength (in kg) of each fibre was tested. The sample mean was $\bar{x} = 150.72$ kg and the sample variance was $s^2 = 37.75$ kg$^2$. A confidence interval for the variance is required.

Assuming a normal model, we might wish to find, say, a 90% confidence interval for the population variance $\sigma^2$. The theoretical distribution underlying statistical inferences about a normal variance is the $\chi^2$-distribution, which as you saw in Chapter 6 is actually a whole family of distributions whose indexing parameter (the degrees of freedom) is, in this particular context, dependent upon the sample size. Here, again, is the sampling distribution of the variance of normal random samples of size $n$, when the parent population is $N(\mu, \sigma^2)$:

$$\frac{(n - 1)s^2}{\sigma^2} \sim \chi^2(n - 1). \tag{7.5}$$

This result was first used in Exercise 6.26.
In Figure 7.3 you can see how the chi-squared density changes shape with increasing degrees of freedom.

![Figure 7.3](image)

**Figure 7.3** The density of $\chi^2(\nu)$, $\nu = 2, 3, 10$

For small values of the parameter $\nu$ the $\chi^2$ density is very highly skewed; for increasing values of the parameter the distribution becomes less skewed.

Corresponding critical points for the chi-squared distribution can be obtained by reference to statistical tables (or by keying in the appropriate command on your computer). For instance, the 5% point of $\chi^2(4)$ is 0.711; the 97.5% point is 11.143. This information is shown graphically in Figure 7.4. You should check that you can obtain these values from statistical tables, and from your computer.

![Figure 7.4](image)

**Figure 7.4** Critical points of $\chi^2(4)$

A confidence interval for an unknown variance $\sigma^2$ based on a random sample $x_1, x_2, \ldots, x_n$ from a normal distribution may be found as follows. As usual, we begin with a probability statement based on the distribution of the estimator (in this case, $S^2$). Using (7.5) we can write

$$P \left( c_L \leq \frac{(n-1)S^2}{\sigma^2} \leq c_U \right) = 1 - \alpha,$$

(7.6)

where $c_L = q_{1-\frac{1}{2}\alpha}$ is the 'left-hand' critical point of the $\chi^2(n-1)$ distribution; and $c_U = q_{1-\frac{1}{2}\alpha}$ is the 'right-hand' critical point.
The double inequality on the left-hand side of (7.6) can be written with the variance $\sigma^2$ as the subject:

$$P\left(\frac{(n-1)s^2}{c_U} \leq \sigma^2 \leq \frac{(n-1)s^2}{c_L}\right) = 1 - \alpha. \quad (7.7)$$

This is a probability statement about a random interval which, with probability $1 - \alpha$, will contain the unknown variance $\sigma^2$. The corresponding confidence interval for $\sigma^2$, based on a random sample with standard deviation $s$, is given by

$$(\sigma^2_-, \sigma^2_+) = \left(\frac{(n-1)s^2}{c_U}, \frac{(n-1)s^2}{c_L}\right). \quad (7.8)$$

**Example 7.8 Finding a confidence interval for a normal variance**

For inferences from normal samples of size 10, say, we would refer to tables of the $\chi^2(9)$ distribution: the 5% point, for example, is 3.325, the 95% point is 16.919. Different samples of size 10 from $N(\mu, \sigma^2)$ will give rise to different observations on the sample variance $S^2$—it is a random variable—and we can write down the probability statement

$$P\left(3.325 \leq \frac{9s^2}{\sigma^2} \leq 16.919\right) = 0.90.$$

The corresponding 90% confidence interval for $\sigma^2$ is

$$(\sigma^2_-, \sigma^2_+) = \left(\frac{(n-1)s^2}{c_U}, \frac{(n-1)s^2}{c_L}\right) = \left(\frac{9s^2}{16.919}, \frac{9s^2}{3.325}\right) = (0.53s^2, 2.71s^2).$$

Notice that in the whole of this example, knowledge of the sample mean is unnecessary. ■

**Example 7.6 continued**

For the kitchen timer data, we have $(n-1)s^2 = 9 \times 3.1232 = 28.109$; it follows that a 90% confidence interval for $\sigma^2$, using $c_L = 3.325$, $c_U = 16.919$, is given by

$$(\sigma^2_-, \sigma^2_+) = \left(\frac{28.109}{16.919}, \frac{28.109}{3.325}\right) = (1.66, 8.45).$$

(This may be compared with the point estimate $s^2 = 3.12$.) ■

**Exercise 7.15**

For the breaking strength data of Example 7.7, find (a) a 90% confidence interval and (b) a 95% confidence interval for $\sigma^2$. 

---

289
7.3.3 Confidence intervals for the normal standard deviation

Returning once more to the probability statement (7.7),

\[ P \left( \frac{(n-1)S^2}{c_U} \leq \sigma^2 \leq \frac{(n-1)S^2}{c_L} \right) = 1 - \alpha, \]

and then taking square roots, we obtain

\[ P \left( S\sqrt{\frac{n-1}{c_U}} \leq \sigma \leq S\sqrt{\frac{n-1}{c_L}} \right) = 1 - \alpha. \]

It follows that a 100(1 - \alpha)% confidence interval for \( \sigma \), based on a random sample with standard deviation \( S \), is given by

\[ (\sigma_-, \sigma_+) = \left( \sqrt{\frac{n-1}{c_U}}, \sqrt{\frac{n-1}{c_L}} \right). \]

In other words, the confidence limits for \( \sigma \) are just the square roots of the respective confidence limits for \( \sigma^2 \).

Exercise 7.16

For the breaking strength data of Example 7.7, (a) estimate \( \sigma \); and (b) find a 90% confidence interval for \( \sigma \). (Use your answers from Exercise 7.15.)

7.4 Larger samples

We have seen that confidence intervals for the parameters of a normal distribution have been fairly easy to develop by following these rules:

(i) write down a probability statement about a sample statistic whose distribution involves the parameter for which the confidence interval is required;

(ii) turn this probability statement into a statement about a random interval which might or might not contain the unknown parameter;

(iii) collect a random sample;

(iv) calculate numerical values for the extremes of the random interval;

(v) call your single observation on that random interval a confidence interval.

In this section we return to non-normal models such as the exponential and Bernoulli distributions, and develop techniques for finding confidence intervals for unknown parameters with the assumption that the random samples we have drawn, in order to perform such inferences, are large enough for the central limit theorem to apply. (Where this assumption cannot be made, we need to use the exact methods as discussed in Section 7.2.)

Essentially, the idea is this: if \( X_1, X_2, \ldots, X_n \) are observations on a random variable \( X \), then the central limit theorem says that

\[ \sum_{i=1}^{n} X_i = X_1 + X_2 + X_3 + \cdots + X_n \approx N(n\mu_X, n\sigma_X^2), \]

where the moments \( \mu_X \) and \( \sigma_X^2 \) are respectively the mean and variance of the random variable \( X \). That is, the sample total is approximately normally
distributed. The moments $\mu_X$ and $\sigma^2_X$ will usually feature the unknown parameter (or parameters, if there is more than one) that we wish to estimate. So this result can be used to write down approximate probability statements about the sample total which can be turned into confidence statements about the unknown parameters in the usual way.

### 7.4.1 The exponential model

One of the simplest applications of this idea is to data for which an exponential distribution is assumed to provide an adequate model.

#### Example 7.9 Coal-mining disasters

Here is the full data set from the investigation of Example 7.2. Table 7.4 gives the times (in days) between consecutive disasters in coal mines in Britain between 15 March 1851 and 22 March 1962. (The data are to be read across rows.) This time interval covers 40,550 days. There were 191 explosions altogether, including one on each of the first and last days of the investigation. So the data involve 190 numbers whose sum is 40,549. The 0 occurs because there were two disasters on 6 December 1875.

<table>
<thead>
<tr>
<th>Table 7.4</th>
<th>Times (in days) between disasters</th>
</tr>
</thead>
<tbody>
<tr>
<td>157</td>
<td>123</td>
</tr>
<tr>
<td>33</td>
<td>66</td>
</tr>
<tr>
<td>186</td>
<td>23</td>
</tr>
<tr>
<td>78</td>
<td>202</td>
</tr>
<tr>
<td>538</td>
<td>187</td>
</tr>
<tr>
<td>3</td>
<td>324</td>
</tr>
<tr>
<td>143</td>
<td>16</td>
</tr>
<tr>
<td>193</td>
<td>134</td>
</tr>
<tr>
<td>378</td>
<td>36</td>
</tr>
<tr>
<td>96</td>
<td>124</td>
</tr>
<tr>
<td>59</td>
<td>61</td>
</tr>
<tr>
<td>108</td>
<td>188</td>
</tr>
<tr>
<td>54</td>
<td>217</td>
</tr>
<tr>
<td>275</td>
<td>78</td>
</tr>
<tr>
<td>498</td>
<td>49</td>
</tr>
<tr>
<td>806</td>
<td>517</td>
</tr>
<tr>
<td>275</td>
<td>20</td>
</tr>
<tr>
<td>330</td>
<td>312</td>
</tr>
<tr>
<td>129</td>
<td>1630</td>
</tr>
</tbody>
</table>

Suppose that these data represent a random sample from an exponential distribution and that we are interested in the mean time interval between disasters. So in this example we have data, a specified model, and we have isolated a parameter of interest.

Suppose we denote the exponential mean by $\mu$. Then denoting by $X_i$ the $i$th element of the sample, we have $\mu_X = \mu$, $\sigma^2_X = \mu^2$, and it follows that

$$E\left(\sum_{i=1}^{n} X_i\right) = n\mu \quad \text{and} \quad V\left(\sum_{i=1}^{n} X_i\right) = n\mu^2.$$  

Applying the central limit theorem,

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\mu\sqrt{n}} = \frac{\bar{X} - \mu}{\mu/\sqrt{n}} \approx N(0,1).$$  

The variance of an exponential distribution is equal to the square of its mean: see (4.25).
Notice the ‘$\approx$’ symbol: the normal distribution is only an approximate model for the sample total when the sample size is large, and we must remember this when making confidence statements. There is no point in giving confidence limits to six significant figures.

Now let us proceed as before to try to find a 95% confidence interval for the exponential mean in the case of the coal-mining data. Notice that the only sample statistic in the expression at (7.9) is the sample mean $\bar{X}$; it is not necessary to calculate the sample standard deviation. Using (7.9), we can write the probability statement

$$P\left(-1.96 \leq \frac{\bar{X} - \mu}{\mu/\sqrt{n}} \leq 1.96\right) \approx 0.95.$$  

This is a statement about the random variable $\bar{X}$. We render it into a statement about a random interval by rearranging the double inequality so that $\mu$ is the subject. By rewriting the inequality in this way, we obtain

$$P\left(\frac{\bar{X}}{1 + 1.96/\sqrt{n}} \leq \mu \leq \frac{\bar{X}}{1 - 1.96/\sqrt{n}}\right) \approx 0.95. \quad (7.10)$$

This says that the random interval defined by transforming the sample mean in the manner described will, with high probability, contain the unknown population mean $\mu$: it provides a 95% confidence interval for $\mu$.

For the coal-mining disaster data, the mean of the $n = 190$ waiting times is $\bar{x} = 213.416$ (days). Based on these data, the corresponding observation on the random interval defined at (7.10) is

$$\left(\mu_{-}, \mu_{+}\right) \approx \left(\frac{\bar{x}}{1 + 1.96/\sqrt{190}}, \frac{\bar{x}}{1 - 1.96/\sqrt{190}}\right)$$

$$= \left(\frac{213.416}{1 + 1.96/\sqrt{190}}, \frac{213.416}{1 - 1.96/\sqrt{190}}\right)$$

$$= \left(\frac{213.416}{1 + 0.14219}, \frac{213.416}{1 - 0.14219}\right)$$

$$= (186.8, 248.8).$$

Since for exponential random samples the maximum likelihood estimate $\hat{\mu}$ of $\mu$ is the sample mean $\bar{x}$, the preceding discussion can be summarized as follows.

**Approximate large-sample confidence interval for an exponential mean**

An approximate $100(1 - \alpha)%$ confidence interval for the mean $\mu$ of an exponential random variable, based on a random sample $x_1, x_2, \ldots, x_n$, is given by

$$\left(\mu_{-}, \mu_{+}\right) \approx \left(\frac{\hat{\mu}}{1 + z/\sqrt{n}}, \frac{\hat{\mu}}{1 - z/\sqrt{n}}\right),$$

where $z$ is the $100(1 - \frac{1}{2}\alpha)%$ point of the standard normal distribution and $\hat{\mu} = \bar{x}$ is the maximum likelihood estimate of $\mu$. 

292
Notice that the fact that these are approximate limits is demonstrated by the presence of the term, \(1 - z/\sqrt{n}\), which if the sample is too small or if the confidence level sought is too high, will be negative. This does not make sense.

**Exercise 7.17**

Use the earthquake data of *Chapter 4*, Table 4.7 to establish a 95% confidence interval for the mean time between serious earthquakes world-wide. State any modelling assumptions you make. (Use your answers to Exercise 7.9.)

**Exercise 7.18**

(a) Establish a 90% confidence interval for the mean traffic rate (vehicles per minute) for the Kwinana Freeway, based on the traffic data of Table 2.11, stating any assumptions you make during the calculation.

(b) Comment on any differences between the confidence interval you have found in part (a) and the one you calculated in Exercise 7.9.

A confidence interval which is wide is less useful than one which is narrow. Large samples give narrower and hence more useful confidence limits than small samples. It is possible to use the general expression for the confidence limits to determine how large the sample should be before the experiment is performed. This is demonstrated in the following example.

**Example 7.10 Determination of the sample size**

Suppose that a sample of observations from an exponential distribution is required to produce a 90% confidence interval for the mean, with both upper and lower confidence limits within 5% of the estimated mean. This implies that both the following inequalities must hold:

\[
\frac{\hat{\mu}}{1 + z/\sqrt{n}} \geq 0.95\hat{\mu} \quad \text{and} \quad \frac{\hat{\mu}}{1 - z/\sqrt{n}} \leq 1.05\hat{\mu},
\]

where \(z = 1.645\). The first inequality gives \(\sqrt{n} \geq 19z\), or \(n \geq 976.9\); the second inequality gives \(\sqrt{n} \geq 21z\), or \(n \geq 1193.4\). So a sample size of at least 1194 (say 1200, perhaps) will ensure a sufficiently narrow confidence interval at the specified 90% level.

**Exercise 7.19**

A sample of observations from an exponential distribution is required. It must be sufficiently large to ensure that a 95% confidence interval for the mean has upper and lower confidence limits both within 3% of the estimated mean. What minimum sample size will achieve this?

**7.4.2 The Poisson model**

Assume that we have a random sample of size \(n\) from a Poisson distribution with unknown mean \(\mu\). Then, by the central limit theorem, the sample total
\[ \sum_{i=1}^{n} X_i \text{ is approximately normally distributed with mean } n\mu \text{ and variance } n\mu. \] For instance, we can make the probability statement

\[ P \left( -1.645 \leq \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\mu}} \leq 1.645 \right) \approx 0.90. \]

In order to calculate a confidence interval for \( \mu \), we need to solve the double inequality inside \( P(\cdot) \) and make \( \mu \) the subject. To generalize what follows to confidence levels other than 90\%, we shall first replace the number 1.645 by \( z \). Then, squaring both sides, we obtain (after a little algebra) the following quadratic inequality in \( \mu \):

\[ n\mu^2 - (2n\bar{X} + z^2)\mu + n\bar{X}^2 \leq 0. \]

The inequality is satisfied for \( \mu \) lying between the two solutions of the quadratic, which are

\[ \frac{2n\bar{X} + z^2 \pm \sqrt{4n\bar{X}z^2 + z^4}}{2n}. \]

This is a very complicated expression that could nevertheless be used to define upper and lower confidence limits for \( \mu \). However, it is based on an approximation (that induced by applying the central limit theorem and using approximate normal theory) and, in the interests of simplifying things a little, a second approximation is now introduced which should not seriously perturb matters further. In the numerator of the expression above, terms involving \( n \) will dominate, since \( n \) is assumed to be large. Including only these dominant terms, the two solutions are approximately

\[ \frac{2n\bar{X} \pm \sqrt{4n\bar{X}z^2}}{2n} = \bar{X} \pm z \sqrt{\frac{\bar{X}}{n}}. \]

Of course, in the Poisson case, the sample mean \( \bar{X} \) is simply the maximum likelihood estimator \( \hat{\mu} \). So, writing these solutions in their most general form, we have the following result.

**Approximate large-sample confidence interval for a Poisson mean**

If a random sample \( x_1, x_2, \ldots, x_n \) is obtained from a population where the Poisson distribution is assumed to provide an adequate model for variation, then an approximate 100(1 - \( \alpha \))% confidence interval for the population mean \( \mu \) is given by

\[ \left( \hat{\mu} - z \sqrt{\frac{\hat{\mu}}{n}}, \hat{\mu} + z \sqrt{\frac{\hat{\mu}}{n}} \right), \]

where \( \hat{\mu} \) is equal to \( \bar{X} \), the sample mean, and \( z \) is the 100(1 - \( \frac{1}{2} \alpha \))% point of the standard normal distribution.
Exercise 7.20

The full childhood accident data initially described in Example 7.1 are given in Table 7.5.

Table 7.5  621 childhood accident counts, aged 4 to 7 and 8 to 11

<table>
<thead>
<tr>
<th>Injuries aged 4 to 7</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>101</td>
<td>76</td>
<td>35</td>
<td>15</td>
<td>7</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>167</td>
<td>61</td>
<td>32</td>
<td>14</td>
<td>12</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>24</td>
<td>36</td>
<td>22</td>
<td>15</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>aged 8</td>
<td>3</td>
<td>10</td>
<td>19</td>
<td>10</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>to 11</td>
<td>4</td>
<td>1</td>
<td>7</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>206</td>
<td>201</td>
<td>107</td>
<td>56</td>
<td>29</td>
<td>12</td>
<td>6</td>
<td>4</td>
</tr>
</tbody>
</table>

The columns give injury counts between the ages of 4 and 7; the rows give the counts between the ages of 8 and 11. For instance, by reference to the column labelled ‘2’ and the row labelled ‘1’, you can see that 32 of the 621 children sampled experienced two accidents between the ages of 4 and 7, followed by just one accident between the ages of 8 and 11.

Summing across diagonals gives the frequencies for the total number of accidents sustained over the complete eight-year period. For example, the total number of children who had two accidents is $24 + 61 + 35 = 120$. The data are summarized in Table 7.6.

Table 7.6  621 childhood accident counts, aged 4 to 11

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>143</td>
<td>120</td>
<td>93</td>
<td>63</td>
<td>49</td>
<td>23</td>
<td>13</td>
<td>12</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Use the data in Table 7.6 to calculate a 95% confidence interval for the average number of accidents sustained by children between the ages of 4 and 11. State any modelling assumptions you make.

A reasonable rule of thumb for applying the central limit theorem to Poisson data in order to calculate approximate confidence intervals for the mean $\mu$, is that the sample total should be at least 30. (So, in this case, notice that you need not worry about the sample size.)

7.4.3 The geometric model

Approximate confidence intervals for the parameter $p$ of a geometric distribution, and for the mean $\mu = 1/p$, are given overleaf.

You are spared the details of the algebra here. The result is important. Notice that the confidence limits for the mean $\mu$ in this case are not the reciprocals of those for the parameter $p$. This is because the limits were obtained in each case through an approximation, and what gets ‘cut’ differs in the two cases.
Approximate large-sample confidence intervals for the parameter of a geometric distribution

An approximate $100(1 - \alpha)\%$ confidence interval for the parameter $p$ of a geometric distribution, based on a random sample $x_1, x_2, \ldots, x_n$ is given by

$$(p_-, p_+) = \left( \hat{p} - z\sqrt{\frac{1 - \hat{p}}{n}}, \hat{p} + z\sqrt{\frac{1 - \hat{p}}{n}} \right),$$

where $\hat{p} = 1/x$ is the reciprocal of the sample mean and $z$ is the $100(1 - \frac{1}{2}\alpha)\%$ point of the standard normal distribution. Equivalently, an approximate $100(1 - \alpha)\%$ confidence interval for the population mean $\mu$ is given by

$$(\mu_-, \mu_+) = \left( \hat{\mu} - z\sqrt{\frac{\hat{\mu}(\hat{\mu} - 1)}{n}}, \hat{\mu} + z\sqrt{\frac{\hat{\mu}(\hat{\mu} - 1)}{n}} \right),$$

where $\hat{\mu}$ is the sample mean $\bar{x}$.

**Exercise 7.21**

Chapter 6, Table 6.5 gives the lengths of runs of Douglas firs infected with *Armillaria* root rot. Use the data to find a 99% confidence interval for the mean length of a run, and state any modelling assumptions you make.

### 7.4.4 The Bernoulli model

Approximate confidence intervals for the Bernoulli parameter, the probability $p$, may be obtained as follows.

Approximate large-sample confidence interval for the Bernoulli probability $p$

An approximate $100(1 - \alpha)\%$ confidence interval for the Bernoulli probability $p$, based on observing $x$ successes in a sequence of $n$ independent Bernoulli trials, is given by

$$(p_-, p_+) = \left( \hat{p} - z\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \hat{p} + z\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right),$$

where $\hat{p} = x/n$ is the maximum likelihood estimate of $p$, and where $z$ is the $100(1 - \frac{1}{2}\alpha)\%$ point of the standard normal distribution.

**Example 7.11 Smokers**

Of a random sample of 7383 adults aged 18 and over chosen from the electoral register in England, Wales and Scotland in part of a study into low blood pressure, a proportion of 32.8% were found to be smokers.

The reference does not give the actual number of smokers in the sample of 7383. From other information supplied, it is possible only to deduce that the number is between 2419 and 2421. But the number is not necessary to calculate the confidence interval. Based on these data, a 95% confidence interval for the unknown proportion \( p \) of smokers in the population is

\[
(p_-, p_+) = \left( \hat{p} - z \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)
\]

\[
= \left( 0.328 - 1.96 \sqrt{\frac{0.328 \times 0.672}{7383}}, 0.328 + 1.96 \sqrt{\frac{0.328 \times 0.672}{7383}} \right)
\]

\[
= (0.328 - 0.011, 0.328 + 0.011)
\]

\[
= (0.317, 0.339).
\]

Notice that in the Bernoulli case the width of the confidence interval is \( 2z \sqrt{\hat{p}(1-\hat{p})/n} \). Now, it is fairly easy to see that for \( 0 \leq \hat{p} \leq 1 \) the maximum value that the product \( \hat{p}(1-\hat{p}) \) can take is \( \frac{1}{4} \), so the maximum possible width of the confidence interval is \( z/\sqrt{n} \). If, a 98% confidence interval for \( p \) (so \( z = 2.326 \)) not wider than 0.05 is required, then in the worst case the sample size will have to be at least 2165.

\[\text{Figure 7.5 gives the graph of the function } \hat{p}(1-\hat{p}) \text{ for } 0 \leq \hat{p} \leq 1. \]

The graph is symmetric about \( \hat{p} = \frac{1}{2} \), where it attains its maximum value of \( \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \).

As long as the number of 1s (Yeses) and the number of 0s (Noes) in your sample are both more than about five or six, then approximate confidence intervals calculated after applying the central limit theorem and using normal critical values should be reasonably accurate.

**Exercise 7.22**

An experiment was undertaken to examine the association between eye and hair colour in samples of Scottish school children.

(a) Of 5387 children sampled in Caithness, 286 had hair classified as 'red'. Use these data to establish a 90% confidence interval for the proportion of Scottish children who have red hair. State any modelling and sampling assumptions you make.

(b) Of 5789 fair-haired children sampled in Aberdeen, 1368 had blue eyes. Use this information to determine a 95% confidence interval for the proportion of fair-haired Scottish children who have blue eyes. State any assumptions you make.

7.5 Inference without a model

In all the cases we have met so far some sort of model for the variation observed has been postulated. This raises the question of how to begin a statistical analysis when no model for the data seems apparent either from physical considerations or from some underlying structure in the 'shape' of the data.

What should your approach be if you are simply presented with a large mass of data, with the request (i) to estimate the mean \( \mu \) of the population from which the data were drawn, and (ii) to construct confidence intervals for the population mean?

In this section we shall only be concerned with confidence intervals for the population mean, and not with other population characteristics.

Probably the first thing you need to assume is that the data do indeed constitute a random sample from the population. (This does not necessarily hold, however, and you should not take this assumption lightly.)

Next, it is probably reasonable to suppose that the best estimate for the unknown population mean \( \mu \) is the sample mean \( \bar{x} \). (This will not necessarily be the maximum likelihood estimate for \( \mu \); but the sample mean \( \bar{x} \) will be unbiased for \( \mu \).)

Then, if the sample size \( n \) is reasonably large, the central limit theorem can be applied, and you can assume that the sample mean \( X \) is approximately normally distributed. That is,

\[
\bar{X} \approx N \left( \mu, \frac{\sigma^2}{n} \right),
\]

where \( \sigma^2 \) is the unknown variance in the population. From this there follows the approximate probability statement

\[
P \left( \mu - z \frac{\sigma}{\sqrt{n}} \leq \bar{X} \leq \mu + z \frac{\sigma}{\sqrt{n}} \right) \approx 1 - \alpha,
\]

where \( z \) is the \( 100(1 - \frac{1}{2}\alpha) \)\% point of the standard normal distribution. This can be rewritten in terms of a probability statement about a random interval:

\[
P \left( \bar{X} - z \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z \frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha. \tag{7.11}
\]

The standard deviation \( \sigma \) is of course unknown. If it was known that the underlying variation in the population was normal, then you would substitute the sample standard deviation \( s \) for \( \sigma \), and the appropriate critical \( t \)-value for \( z \). Here, the obvious thing to do first is to replace the unknown standard deviation \( \sigma \) by its estimate \( s \). (There are arguments that \( \sigma \) might just as well be replaced by the maximum likelihood estimate \( \hat{\sigma} \) with the \( n \)-divisor as by \( s \); but these are not necessarily very good arguments.)

There is, however, no good reason for replacing \( z \) in (7.11) by a critical \( t \)-value, even if the sample size is not great: the development of Student's \( t \)-distribution was as an accurate mathematical 'formula' for making probability statements about the distribution of the sample mean \( \bar{X} \) when the
parent population is known to be normal, which is not the case here. So, in the end, what happens is quite simple: we just substitute $s$ for $\sigma$ and leave $z$ as it is. This leads to the following rule.

For a random sample $x_1, x_2, \ldots, x_n$ drawn from a population where no distributional assumptions can be made, an approximate $100(1 - \alpha)\%$ confidence interval for the population mean $\mu$ is given by

$$
(\mu_-, \mu_+) = \left( \bar{x} - z \frac{s}{\sqrt{n}}, \bar{x} + z \frac{s}{\sqrt{n}} \right),
$$

where $n$ is the sample size, $\bar{x}$ is the sample mean, $s$ is the sample standard deviation and $z$ is the $100(1 - \frac{1}{2}\alpha)\%$ point of the standard normal distribution.

**Exercise 7.23**

Chapter 2, Table 2.9 gives data on library book usage: the number of times that each of 122 books was borrowed during the course of a year was counted.

Without making any distributional assumptions about a statistical model for the variation in the number of annual withdrawals, use the data in Table 2.9 to calculate a 90% confidence interval for the mean number of loans in a year.

**Exercise 7.24**

An experiment was conducted into the effects of environmental pollutants upon animals. For 65 Anacapa pelican eggs, the concentration, in parts per million of PCB (polychlorinated biphenyl, an industrial pollutant), was measured, along with the thickness of the shell in millimetres.

The data are summarized in Table 7.7.

<table>
<thead>
<tr>
<th>ppm</th>
<th>mm</th>
<th>ppm</th>
<th>mm</th>
<th>ppm</th>
<th>mm</th>
<th>ppm</th>
<th>mm</th>
</tr>
</thead>
<tbody>
<tr>
<td>452</td>
<td>0.14</td>
<td>184</td>
<td>0.19</td>
<td>115</td>
<td>0.20</td>
<td>315</td>
<td>0.20</td>
</tr>
<tr>
<td>139</td>
<td>0.21</td>
<td>177</td>
<td>0.22</td>
<td>214</td>
<td>0.22</td>
<td>356</td>
<td>0.22</td>
</tr>
<tr>
<td>166</td>
<td>0.23</td>
<td>246</td>
<td>0.23</td>
<td>177</td>
<td>0.23</td>
<td>289</td>
<td>0.23</td>
</tr>
<tr>
<td>175</td>
<td>0.24</td>
<td>296</td>
<td>0.25</td>
<td>205</td>
<td>0.25</td>
<td>324</td>
<td>0.26</td>
</tr>
<tr>
<td>260</td>
<td>0.26</td>
<td>188</td>
<td>0.26</td>
<td>208</td>
<td>0.26</td>
<td>109</td>
<td>0.27</td>
</tr>
<tr>
<td>204</td>
<td>0.28</td>
<td>89</td>
<td>0.28</td>
<td>320</td>
<td>0.28</td>
<td>265</td>
<td>0.29</td>
</tr>
<tr>
<td>138</td>
<td>0.29</td>
<td>198</td>
<td>0.29</td>
<td>191</td>
<td>0.29</td>
<td>193</td>
<td>0.29</td>
</tr>
<tr>
<td>316</td>
<td>0.29</td>
<td>122</td>
<td>0.30</td>
<td>305</td>
<td>0.30</td>
<td>203</td>
<td>0.30</td>
</tr>
<tr>
<td>396</td>
<td>0.30</td>
<td>250</td>
<td>0.30</td>
<td>230</td>
<td>0.30</td>
<td>214</td>
<td>0.30</td>
</tr>
<tr>
<td>46</td>
<td>0.31</td>
<td>256</td>
<td>0.31</td>
<td>204</td>
<td>0.32</td>
<td>150</td>
<td>0.34</td>
</tr>
<tr>
<td>218</td>
<td>0.34</td>
<td>261</td>
<td>0.34</td>
<td>143</td>
<td>0.35</td>
<td>229</td>
<td>0.35</td>
</tr>
<tr>
<td>173</td>
<td>0.36</td>
<td>132</td>
<td>0.36</td>
<td>175</td>
<td>0.36</td>
<td>236</td>
<td>0.37</td>
</tr>
<tr>
<td>220</td>
<td>0.37</td>
<td>212</td>
<td>0.37</td>
<td>119</td>
<td>0.39</td>
<td>144</td>
<td>0.39</td>
</tr>
<tr>
<td>147</td>
<td>0.39</td>
<td>171</td>
<td>0.40</td>
<td>216</td>
<td>0.41</td>
<td>232</td>
<td>0.41</td>
</tr>
<tr>
<td>216</td>
<td>0.42</td>
<td>164</td>
<td>0.42</td>
<td>185</td>
<td>0.42</td>
<td>87</td>
<td>0.44</td>
</tr>
<tr>
<td>216</td>
<td>0.46</td>
<td>199</td>
<td>0.46</td>
<td>236</td>
<td>0.47</td>
<td>237</td>
<td>0.49</td>
</tr>
<tr>
<td>206</td>
<td>0.49</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We shall discover methods for exploring the association between shell thickness (or thinness) and degree of contamination later in the course. For the moment, use the data on shell thickness to determine a 95% confidence interval for the mean thickness of Anacapa pelican eggs.

**Summary**

1. If a single observation $x$ is taken on a random variable $X$ with a stated probability distribution indexed by a parameter $\theta$ whose value is unknown, then a $100(1 - \alpha)\%$ confidence interval $(\theta_-, \theta_+)$ may be found by solving separately for $\theta$ the two equations

   \[ P(X \leq x) = \frac{1}{2} \alpha \quad \text{and} \quad P(X \geq x) = \frac{1}{2} \alpha. \]

   The two solutions $\theta_-$ and $\theta_+$ are called respectively the lower and upper confidence limits for the parameter $\theta$.

2. The interpretation of a confidence interval is as follows: in independent repeated experiments each resulting in the statement of a $100(1 - \alpha)\%$ confidence interval for an unknown parameter $\theta$, the expected proportion of intervals that actually contain the number $\theta$ is $1 - \alpha$.

3. For small samples from populations where the standard models are assumed, exact confidence limits may be calculated by reference to a statistical computer package; occasionally exact arithmetic computations are possible.

4. For a random sample $X_1, X_2, \ldots, X_n$ of size $n$ from the normal distribution with unknown mean $\mu$ and unknown standard deviation $\sigma$, the sampling distributions of the sample mean $\bar{X}$ and the sample variance $S^2$ are as follows.

   The random variable

   \[ T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \]

   is said to follow Student's $t$-distribution with $n - 1$ degrees of freedom. This is written $T \sim t(n - 1)$. Also,

   \[ \frac{(n - 1)S^2}{\sigma^2} \sim \chi^2(n - 1). \]

   These distributions may be used to calculate confidence intervals for the parameters $\mu$ and $\sigma^2$ (or $\sigma$).
5. For larger samples from standard distributions, the central limit theorem may be applied to construct approximate confidence intervals for unknown parameters. (Occasionally, further approximations are also introduced.)

An approximate large-sample confidence interval for an exponential mean $\mu$ is
\[
(\hat{\mu}_-, \hat{\mu}_+) = \left( \frac{\hat{\mu}}{1 + \frac{z}{\sqrt{n}}}, \frac{\hat{\mu}}{1 - \frac{z}{\sqrt{n}}} \right).
\]

An approximate large-sample confidence interval for a Poisson mean $\mu$ is
\[
\left( \hat{\mu} - z \sqrt{\frac{\hat{\mu}}{n}}, \hat{\mu} + z \sqrt{\frac{\hat{\mu}}{n}} \right).
\]

An approximate large-sample confidence interval for the parameter $p$ in a geometric distribution is given by
\[
(p_-, p_+) = \left( \hat{p} - z\hat{p}\sqrt{\frac{1 - \hat{p}}{n}}, \hat{p} + z\hat{p}\sqrt{\frac{1 - \hat{p}}{n}} \right).
\]

An approximate large-sample confidence interval for the mean $\mu$ of a geometric distribution is given by
\[
(\mu_-, \mu_+) = \left( \hat{\mu} - z \sqrt{\frac{\hat{\mu}(\hat{\mu} - 1)}{n}}, \hat{\mu} + z \sqrt{\frac{\hat{\mu}(\hat{\mu} - 1)}{n}} \right).
\]

An approximate large-sample confidence interval for the Bernoulli probability $p$ is given by
\[
(p_-, p_+) = \left( \hat{p} - z \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \hat{p} + z \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right).
\]

6. An approximate $100(1 - \alpha)\%$ confidence interval for the mean $\mu$ of a population where no distributional assumptions have been made may be obtained by using
\[
(\mu_-, \mu_+) = \left( \bar{x} - z \frac{s}{\sqrt{n}}, \bar{x} + z \frac{s}{\sqrt{n}} \right),
\]
where $n$ is the sample size, $\bar{x}$ is the sample mean, $s$ is the sample standard deviation and $z$ is the $100(1 - \frac{1}{2}\alpha)\%$ point of the standard normal distribution; the sample taken is assumed to be sufficiently large that the central limit theorem may be applied.